DETERMINATION OF CANCER TISSUE HETEROGENEITY

A Thesis

by

ASHISH KATIYAR

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Aniruddha Datta
Committee Members,	P.R. Kumar
	Ulisses Braga Neto
	Sing-Hoi Sze
Head of Department,	Miroslav M. Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

Understanding the heterogeneous nature of cancer tissue is a very important problem in cancer research. It can give insights into the cause of disease, its progression and explain induced drug resistance. There are two models that are used to explain heterogeneity, Cancer Stem Cells and Clonal Evolution. This thesis aims to address this challenge by developing an algorithm to determine the ratio of different components of a heterogeneous cancer tissue. This algorithm is robust and does not depend on the heterogeneity model. The proposed algorithm uses response vector, which is a vector of observable response of cell lines. A database of the response of individual cell lines is developed by collecting cell-by-cell response measurements. A heterogeneous cancer tissue is modeled as being a mixture of these cell lines. Avoiding the high cost cell-by-cell analysis, the collective response of the heterogeneous cancer tissue is observed. The algorithm uses Bayesian inference to estimate the probability distribution of the number of cells of individual cell lines based on the response of individual cell lines and the observed collective response. The results of the algorithm are validated using synthetic data and real-world data collected from cell lines, which are mixed in a ratio known a priori.

ACKNOWLEDGMENTS

I would like to thank Dr. Aniruddha Datta for his motivation and support throughout the research. I would also like to thank Dr. Anwoy Mohanty for providing valuable input.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Aniruddha Datta, Dr. P. R. Kumar and Dr. Ulisses Braga-Neto of the Department of Electrical and Computer Engineering and Dr. Sing-Hoi Sze of the Department of Computer Science and Engineering. The experimental results in Section 3 were provided by Dr. Chao Sima, Jianping Hua and Rosana Lopes. All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by funds from NSF.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND CURRENT STATE OF KNOWLEDGE

1.1 Introduction

Cancer is a disease caused by loss of cell-cycle control. Loss of cell-cycle control results in unregulated cell proliferation and/or reduced cell apoptosis. It may occur due to mutations in proto-oncogenes, which are responsible for the regulation of cell-growth and differentiation, or in tumor suppressor genes that inhibit cell division and survival. Cancer progression can be modeled using two models - Cancer Stem Cells and Clonal Evolution. Both these models have different explanations to account for the heterogeneity observed in cancer tissues.

Diving into cancer tissue heterogeneity is a very crucial factor in cancer treatment. Tracking the composition of cancer tissue can give an insight into the cause and progression of a particular cancer. Heterogeneity has posed a serious challenge in the design of effective therapy. The only kinds of cancers for which a high rate of therapeutic success has been achieved, namely chronic myelogenous leukemia (CML) and acute promyelocytic leukemia (APML), are normally not heterogeneous in nature. Moreover, as a particular dominant subpopulation is targeted for killing via drugs, other subpopulations usually emerge requiring a different therapy. Clearly, this can contribute to the mechanism of acquired drug resistance, which is quite commonly encountered in cancer treatment [1]. Thus, the problem of identifying dominant subpopulations in a cancer tissue is of utmost importance for therapeutic purposes and the goal of the proposed research is to demonstrate the viability of a possible solution to it, based on imaging the fluorescence.

1.2 Current State of Knowledge

There are two models used to explain heterogeneity in cancer tissues:

1.2.1 Cancer Stem Cells

This model asserts that only a small fraction of all the cancer cells, called cancer stem cells, are tumorigenic, that is , they are the ones responsible for the progression of cancer when they reproduce. Their role can be compared to normal stem cells which are responsible for sustaining the tissues and organs. By this theory, the cancer cells which are not stem cells are harmful, but are incapable of sustaining the cancer over a period of time. The proponents of this theory suggest therapies targeting stem cells. The Cancer Stem Cell model explains cancer tissue heterogeneity as arising from differences in the stem cells that give rise to the tumor. This variability in stem cells can arise due to epigenetic changes.

1.2.2 Clonal Evolution

According to this model, cancer arises from mutations in a single cell. This mutation is such that the mutated cell has a proliferative advantage over normal neighboring cells. As this cell multiplies, new mutations accumulate at different cell division steps. Hence as the cancer progresses, new varieties of cancer subpopulations become a part of the cancer tissue making the tissue heterogeneous in nature [2].

A mathematical model to determine the heterogeneity of a cancer tissue is crucial irrespective of the underlying model. It is important for the mathematical model to be independent of the model used to explain heterogeneity as currently there is no consensus on which is more accurate.

There have been various models suggested for determining heterogeneity in a cancer tissue. A model based on gene expression was suggested in [3]. This model represents heterogeneity as a combination of deterministic Boolean Networks based on prior pathway knowledge. It models heterogeneity by representing it as an inherent dirichlet distribution and uses hierarchical Bayesian model to arrive at the parameters of this distribution

from the observed gene expression values. A major drawback of the approach is that the prior knowledge in Boolean Networks is not accurate, which hampers the accuracy of the results. In general, ensemble methods such as these are inaccurate when it comes to determining the ratio of different cell lines. More accurate method for determining the composition of heterogeneous cancer tissue was suggested in [4]. This model observes responses of individual cells and infers which cell line they belong to. This method of inferring the ratio of the components of the heterogeneous tissue is very accurate. However, the high cost of the method makes it unaffordable. In this thesis, we propose an algorithm which overcomes the accuracy constrains of ensemble methods as well as avoids the high cost of a cell by cell analysis method.

2. ALGORITHM

2.1 The Idea

In this thesis we propose an algorithm that overcomes the challenges of ensemble approaches, as well as avoids the high cost of cell-by-cell analysis. We utilize the cell-by-cell analysis to gather the characteristics of the response of individual cell line classes. We can make a database of the characteristics of different cell line classes. This is an expensive process, but it needs to be done only once It is these characteristics that we utilize while performing the analysis of heterogeneous tissue using ensemble methods. That is, once the expensive cell-by-cell analysis is done for individual cell line classes, we can study any number of heterogeneous tissues made up of members of those classes using low cost ensemble methods.

2.2 Model

Let $C = (C_1, C_2, \ldots, C_n)$ be the different types of cells lines in the tissue. Let $\pi = (\pi_1, \pi_2, \ldots \pi_n)$ be the ratio of the corresponding cell lines. Suppose the response of a cell from the i^{th} cell line is given by the random response vector $E_i = (E_{i1}, E_{i2}, \ldots, E_{im})$. This response vector can be quantitatively measured (for example, using invisible or visible fluorescence). The different components of E_i are assumed to be independent. This is a reasonable assumption if the different components represent unrelated quantities. The algorithm can be divided into two steps. The first step is the determination of the parameters of E_i. The second step is the Bayesian estimation of π.

2.2.1 Parameters of E_i

For the determination of parameters of E_i, we utilize the cell-by-cell observation. The objective is to estimate the mean μ_{ij} and standard deviation σ_{ij} of E_{ij} for $1 \leq i \leq n$ and

$1 \leq j \leq m$. The cell-by-cell observations act as samples of the random vector E_i. The sample mean and sample standard deviation represented as $\hat{\mu_{ij}}$ and $\hat{\sigma_{ij}}$ act as estimates of true mean and true standard deviation of E_{ij}. Suppose there are p samples of each cell line. The sample mean $\hat{\mu_{ij}}$ and sample standard deviation $\hat{\sigma_{ij}}$ is given by:

$$\hat{\mu_{ij}} = \frac{1}{n} \sum_{k=1}^{p} e_{ijk} \tag{2.1}$$

$$\hat{\sigma_{ij}} = \sqrt{\frac{\sum_{k=1}^{p} (e_{ijk} - \hat{\mu_{ij}})^2}{n-1}} \tag{2.2}$$

where e_{ijk} are the samples of the random variable E_{ij}.

2.2.2 Bayesian Estimate of π

Obtaining the accurate estimate of π, represented as $\hat{\pi}$, is the objective of the algorithm. In order to achieve this, we proceed with estimating $N = (N_1, N_2, \ldots, N_n)$, where N_i is the number of cells of cell line C_i. $\hat{\pi}$ can be easily calculated as

$$\hat{\pi} = \frac{N}{\sum_{i=1}^{n} N_i} \tag{2.3}$$

The estimation of π utilizes the observation from an ensemble experiment corresponding to the cell-by-cell experiment. As an input, we have the overall response vector, E_{sum} of the heterogeneous cancer tissue. E_{sum} is the result of a summation of the response vector of each cell in the heterogeneous tissue. Each component of E_{sum}, represented by E_{sumj}, is given as:

$$E_{sumj} = E_{1jN1_{sum}} + \ldots + E_{njNn_{sum}} \tag{2.4}$$

where,

$$E_{ijNi_{sum}} = E_{ij1} + \ldots + E_{ijN_i} \tag{2.5}$$

Here, E_{ijk} are independent identically distributed and have the same probability density function as E_{ij} for $1 \leq k \leq N_i$.

Individually, $E_{ijNi_{sum}}$ can be approximated as a Gaussian with mean $N_i\mu_{ij}$ and variance $N_i\sigma_{ij}^2$ by the Central Limit Theorem for sufficiently large N_i. For practical purposes, the cell lines, which form a significant part of heterogeneous cancer tissue, satisfy the condition of large N_i. Hence, the exact distribution of E_{ij} becomes unimportant. This is a very important implication as it gives the independence of choosing any feature as a part of the observation vector irrespective of the probability distribution of the same. The only condition is that the observation of the ensemble should be given by the summation of the observation of individual cells in the tissue.

The likelihood of E_{sumj} can be approximated by:

$$P(E_{sumj}|N,\mu,\sigma) \approx \frac{1}{\sqrt{2\pi(\sum_{i=1}^{n} N_i\sigma_{ij}^2)}} e^{-\frac{(E_{sumj}-\sum_{i=1}^{n}\mu_{ij}Ni)^2}{2\sum_{i=1}^{n}N_i\sigma_{ij}^2}} \tag{2.6}$$

As the components of E_{sum} are independent, the likelihood of E_{sum} is given by:

$$P(E_{sum}|N,\mu,\sigma) = \prod_{j=1}^{m} P(E_{sumj}|N,\mu,\sigma) \tag{2.7}$$

This needs to be maximized over N in order to obtain a maximum likelihood estimate of N. However, the complex expression makes it difficult to solve this problem analytically. Another approach can be to evaluate the expression in (7) for different possible values of N. However, the complexity of the algorithm is exponential and hence it is not feasible when the number of different cell lines is large. Hence, we use a Bayesian approach to estimate N.

The components of N have a uniform prior distribution. All the components of N

are assumed to have a uniform prior from 0 to an arbitrarily large number, say M. The posterior probability of N_i is given by:

$$P(N_i|E_{sum}, N_{-i}, \mu, \sigma) = \frac{P(E_{sum}|N, \mu, \sigma)P(N_i|N_{-i}, \mu, \sigma)}{\int P(E_{sum}|N_i', N_{-i}, \mu, \sigma)P(N_i'|N_{-i}, \mu, \sigma)dN_i'} \tag{2.8}$$

where N_{-i} represents all the components of N excluding the i^{th} component. As N_i is independent of N_{-i}, μ and σ, we have

$$P(N_i|N_{-i}, \mu, \sigma) = P(N_i) = 1/M \tag{2.9}$$

$P(E_{sum}|N, \mu, \sigma)$ can be calculated from 2.7. However, evaluating the denominator term of 2.8 is a complex problem. This makes the problem of calculating the posterior probability of N_i from 2.8 infeasible. To address this issue, we resort to the Metropolis algorithm, which is a Markov chain simulation to estimate the posterior distribution.

2.2.2.1 Metropolis Algorithm

The Metropolis algorithm comes in handy when it is difficult to exactly evaluate the posterior probability. In such a scenario, if it is possible to sample directly from the posterior distribution, we can generate independent identically distributed samples and use them to approximate the posterior probability distribution. However, in our case, it is not possible to sample directly from 2.8. To circumvent this issue, we use the full conditional of N_i which is given by

$$P(N_i|E_{sum}, N_{-i}, \mu, \sigma) \propto P(E_{sum}|N, \mu, \sigma)P(N_i) \tag{2.10}$$

Suppose we have s samples of N_i from the posterior distribution in the set $(N_{i1}, \ldots, N_{is})$. We then consider adding the proposal value N_i^*, which is in the vicinity of N_{is}. We follow the following steps:

1. N_i^* can be obtained by taking a sample from a symmetric proposal distribution. For eg, N_i^* can be sampled from $uniform(N_{is} - \delta, N_{is} + \delta)$.

2. Compute the acceptance ratio $r = P(N_i^*|E_{sum}, N_{-i}, \mu, \sigma)/P(N_{is}|E_{sum}, N_{-i}, \mu, \sigma)$

3. Assign $N_{i(s+1)} = N_i^*$ with probability $min(r, 1)$ or N_{is} otherwise

Substituting $P(E_{sum}|N, \mu, \sigma)$ and $P(N_i)$ from 2.7 and 2.9 in 2.10 while performing step 2, we see that M cancels and hence the algorithm is independent of M. The Markov chain formed by following the aforementioned steps has the same stationary distribution as the posterior distribution of N. The Markov chain needs to run for a few initial iterations before it reaches stationarity and only after that, the sampling has to be done. An important consideration is the length of the neighborhood for the proposal distribution. If the neighborhood is too small, the Markov chain will take too long to reach stationarity and the samples will generate too many samples close to each other. Too large a neighborhood would result in too many samples being rejected once the Markov chain has reached stationarity. Hence the value of neighborhood parameter needs to be tuned appropriately. We draw samples from this Markov Chain after running it till it reaches stationarity. These samples are used to estimate the posterior distribution of N. To do this, we use a non parametric probability density function estimation, Kernel density estimation.

2.2.2.2 Kernel Density Estimation

Let $(N_{i1}, N_{i2}, \ldots, N_{ik})$ be the samples of the posterior distribution of N_i drawn from the Metropolis algorithm. The Kernel Density Estimate of the posterior distribution is given by:

$$\hat{f}_{N_i}(n_i|E_{sum}, N_{-i}, \mu, \sigma) = \frac{1}{kh} \sum_{j=1}^{k} K(\frac{n_i - N_{ij}}{h}) \tag{2.11}$$

Here, K is the Kernel function. Usually, K is a non-negative function with mean 0 and it integrates to 1. In our case, we will consider K to be standard normal.

If K is smooth, the density estimate obtained is also smooth which is the advantage offered by this density estimation method. An important consideration for the accuracy of density estimation is the value of the bandwidth parameter, h. A low value of h results in high variance in the estimation. A high value of h results in high bias in the estimation. The optimal value of h which minimizes the squared error cost is given by:

$$h_{opt} = Dk^{-1/5} \tag{2.12}$$

where $D = \frac{R(K)^{1/5}}{(R(f'')\sigma_K^4)^(1/5)}$ where $R(g) = \int g^2(x)dx$ Since it involves f, where f is the true posterior distribution, it is not possible to calculate the exact value of h. An approximation for the optimal value of h can be obtained assuming f to be Gaussian. This bandwidth is called the plug in bandwidth and is given by the expression

$$\hat{h}_{plugin} = 1.06sk^{-1/5}, s^2 = \frac{1}{k-1}\sum_{j=1}^{k}(N_{ij} - \bar{N}_i)^2 \tag{2.13}$$

Once the posterior density function estimation is done, we can evaluate the posterior mean, the value of N which has maximum posterior probability, the confidence interval, etc. to come to conclusions about the composition of the heterogeneous cancer tissue.

3. RESULTS

The performance of the algorithm is tested for simulated data and experimental data. We use simulated data to study the performance of the algorithm as various parameters change. The parameters taken into account are class of probability distribution of the response vector, similarity of the expression profile, and standard deviation of the components of the response vector. We study their impact on the confidence interval and accuracy of the results.

3.1 Simulated Data

In order to demonstrate the performance of the proposed algorithm, we test its performance on synthetic data. To do so, we first generate the samples for cell-by-cell analysis. We consider a two cell type system. The number of cells of i^{th} type with maximum posterior probability is considered to be the value of N_i. Once N_i has been determined, we can calculate $\hat{\pi}$ using 2.3. The root square error, e, of the estimate of π is used as the parameter to evaluate the performance.

$$e = \sqrt{\sum_{i=1}^{n} (\pi_i - \hat{\pi})^2} \tag{3.1}$$

We use the simulated data to perform a variety of analysis. We first evaluate the algorithm in a scenario where the elements of the response vector are exponentially distributed. In order to perform the analysis, we generate 2000 samples for cell-by-cell analysis to estimate the mean and standard deviation of the individual components in the response vector of the cell lines. The response vector has 2 components. In our analysis we set $\mu_{11} = \mu_{22} = 100$ and $\mu_{12} = \mu_{21} = k\mu_{11}$ where $0 < k < 1$. Changing the value of k varies the similarity between the response profile of the individual cell lines. $k = 0$ corresponds to a case when the response profiles of the individual cell lines are the most dissimilar. $k = 1$ corresponds

to the case when the response profile of both the cell lines are exactly the same and therefore it is not possible to differentiate between the cell lines. Since the inherent distribution is exponential, we have $\sigma_{11} = \sigma_{22} = 100$ and $\sigma_{12} = \sigma_{21} = k\sigma_{11}$. The simulation is performed for 3 different values of k, $k = 0, k = 0.3, k = 0.7$. The idea is to show that as the similarity between the expression profile of the different components decreases, the confidence interval of the posterior probability distribution becomes narrower, that is, the performance of the algorithm improves. We generate the ensemble observation data by taking samples from the same μ and their summation acts as an input to the algorithm. The samples are generated such that Cell Type 1 has 2000 cells and Cell Type 2 has 3000 cells in the heterogeneous mixture. This corresponds to $\pi = (0.4, 0.6)$. This is the case that has been used in all the simulations with synthetic data. While running the Metropolis Algorithm, we let the Markov chain run for 10000 iterations before sampling so that it reaches stationarity. The neighborhood in the Metropolis Algorithm is tuned such that approximately one third of the proposed values are accepted. This procedure is followed for both, the simulated data and the experimental data. The results in Figure 3.1 show that the observation is on the expected lines as we can see the confidence interval getting narrower as the values of k is increasing. The error e for all the cases remains below 0.0060.

Next, we do the same analysis for the scenario when the elements of the response vector are normally distributed. The value of μ remains the same as the exponential case and so does the range of values of k. We assume constant coefficient of variation, which implies that the variance of the normally distributed elements of response vector is a constant multiple of its mean, that is, $\sigma_{ij}^2 = c\mu_{ij}$. The analysis is done for $c = 6$ and we again observe the effect of similarity in expression profiles for the two cell types on the confidence interval. The results as shown in Figure 3.2 shows the confidence interval becoming narrower as the value of k decreases. We also study the error performance as the value of k increases. We observe that after a certain threshold close to 1, the error shoots up as

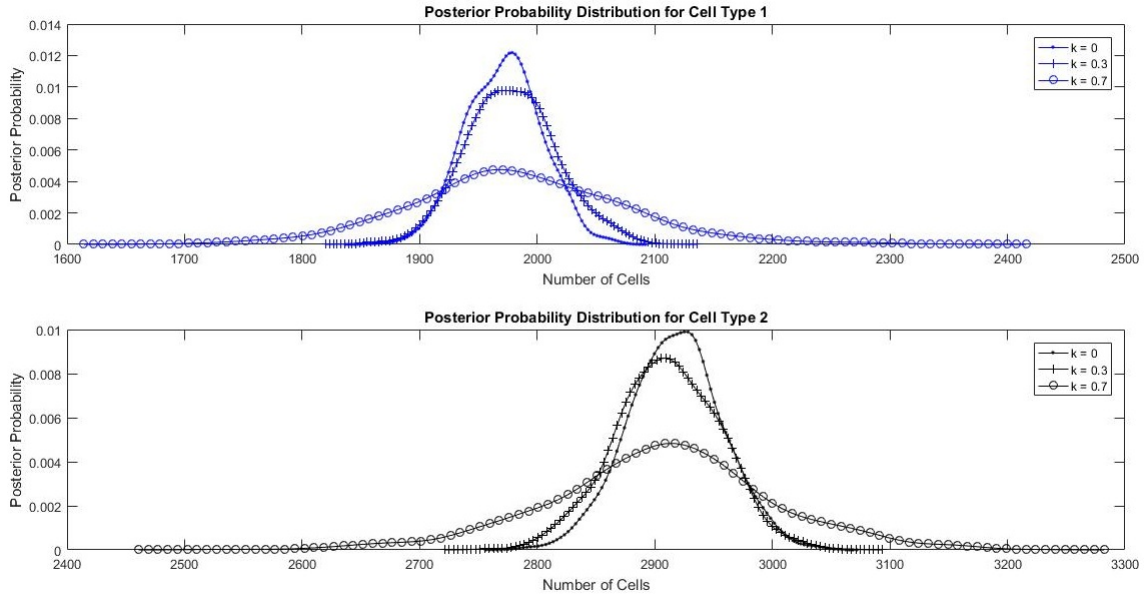

Figure 3.1: Posterior Probability Distribution of N for Exponentially Distributed Elements of the Response Vector

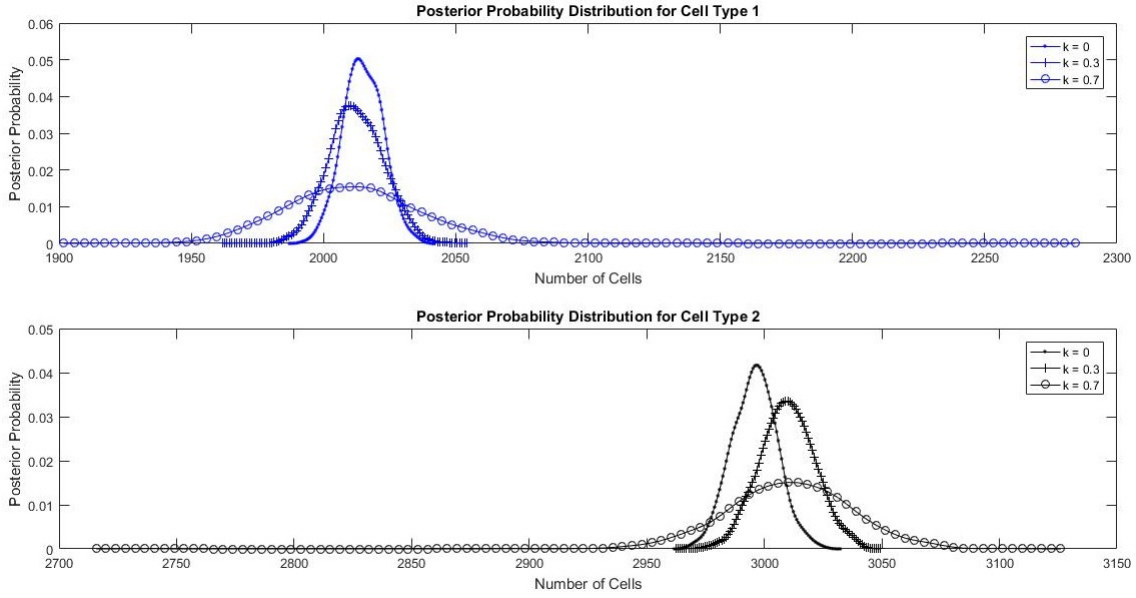

Figure 3.2: Posterior Probability Distribution of N for Gaussian Distributed Elements of the Response Vector

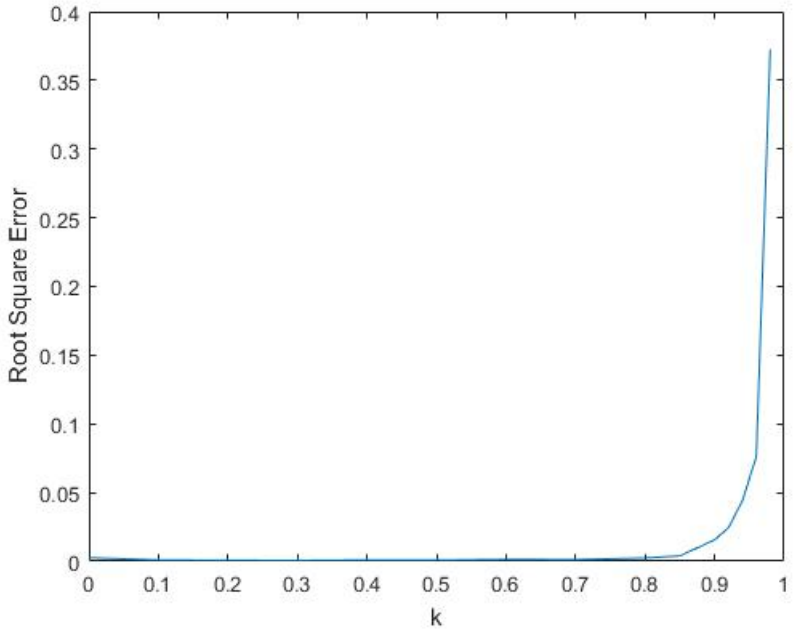

Figure 3.3: Error Performance for Increasing k

demonstrated in Figure 3.3.

We next analyze the impact of the standard deviation of the components of the response vector on the performance of the algorithm. To perform this analysis, we set $\mu_{11} = \mu_{22} = 100$ and $\mu_{12} = \mu_{21} = 0$. To study the effect of the variance, we vary the value of c. For the case when the variance is lower, we expect the confidence interval to be narrower. The results, given in Figure 3.4, show that this is in fact the case upon running the algorithm for $c = 0.5, c = 2, c = 5$ and $c = 10$. We also demonstrate the error performance with increasing value of c, which is shown in Figure 3.5.

3.2 Experimental Data

The algorithm is tested for a heterogeneous mixture of three separate human cancer cell lines, HCT116 (Colon Cancer), A2058(Melanoma) and SW480 (Colorectal carcinoma). The response vector is composed of Red, Green and Blue fluorescence. There are three

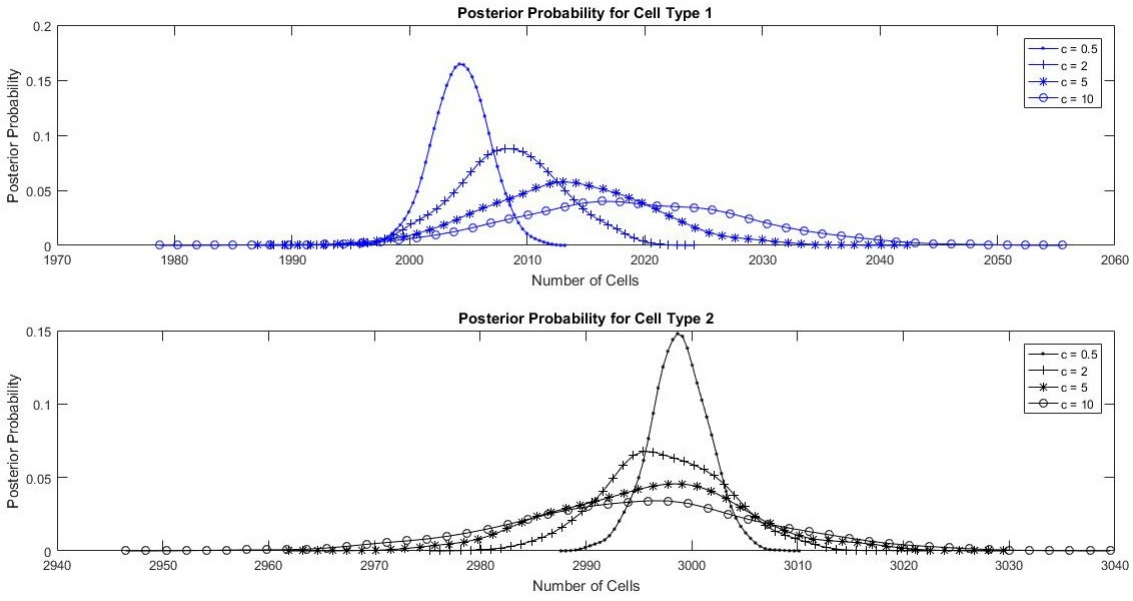

Figure 3.4: Posterior Probability Distribution of N for Increasing Variance

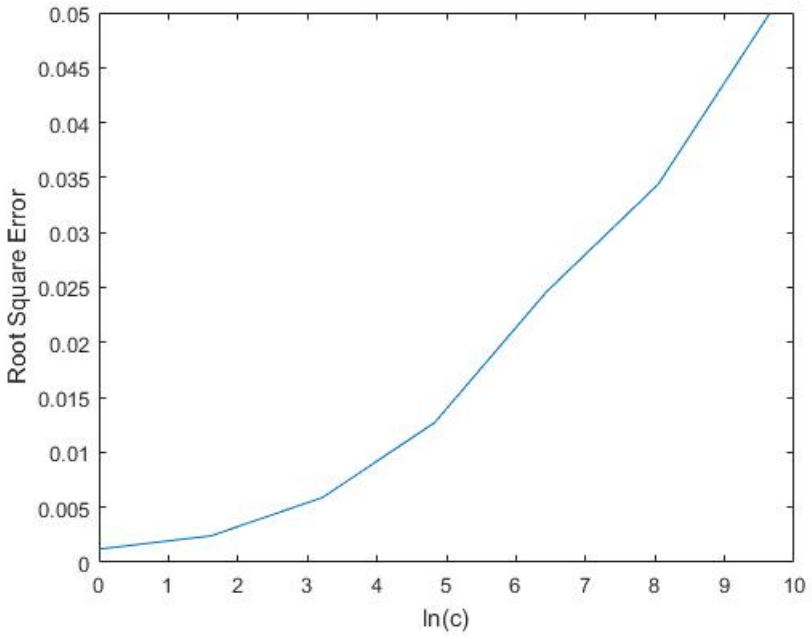

Figure 3.5: Error Performance for Increasing c

Experiment	π	$\hat{\pi}$	e
Untreated Mixture 1	[0.33 0.40 0.27]	[0.40 0.48 0.12]	0.1838
Untreated Mixture 2	[0.37 0.19 0.44]	[0.29 0.16 0.55]	0.1393
Lapatinib Mixture 1	[0.29 0.41 0.30]	[0.24 0.40 0.36]	0.0787
Lapatinib Mixture 2	[0.35 0.16 0.49]	[0.25 0.12 0.63]	0.1766
Temsirolimus Mixture 1	[0.31 0.42 0.28]	[0.32 0.50 0.18]	0.1285
Temsirolimus Mixture 2	[0.41 0.16 0.43]	[0.28 0.14 0.58]	0.1995

Table 3.1: Inherent and Estimated Values of π

cases - the mixture is untreated, treated with Lapatinib and treated with Temsirolimus. For each of the three cases, there are two different mixtures. Hence, overall there are six test cases. In each case we know the true ratio π of the cell lines in the mixture. We evaluate performance by evaluating e as given by Table 3.1. The posterior probability distribution for all the test cases are presented in Figures 3.6, 3.7, 3.8, 3.9, 3.10 and 3.11. The values $\pi, \hat{\pi}$ and e are shown in the table below. The first, second and third components of π correspond to the ratio of HCT116, A2058 and SW480 respectively. As is clear from the observation, in every case, the algorithm determines the concentration of the three cell lines quite well.

3.3 Important Considerations

There are multiple factors crucial for the performance of the algorithm. Firstly, it should be made sure that in the cell-by-cell analysis, enough samples are there to arrive at an accurate estimate of the mean and variance. Secondly, the experimental setup has to be standard as a variation of the setup from the one used for estimation of parameters of response vectors can lead to a change in the parameters when the heterogeneous tissue is being analyzed and will ultimately result in inaccurate results. This phenomena is observed in the experimental results, which explains the accuracy of the algorithm being less for experimental data compared to simulated data. Thirdly, the algorithm performs best

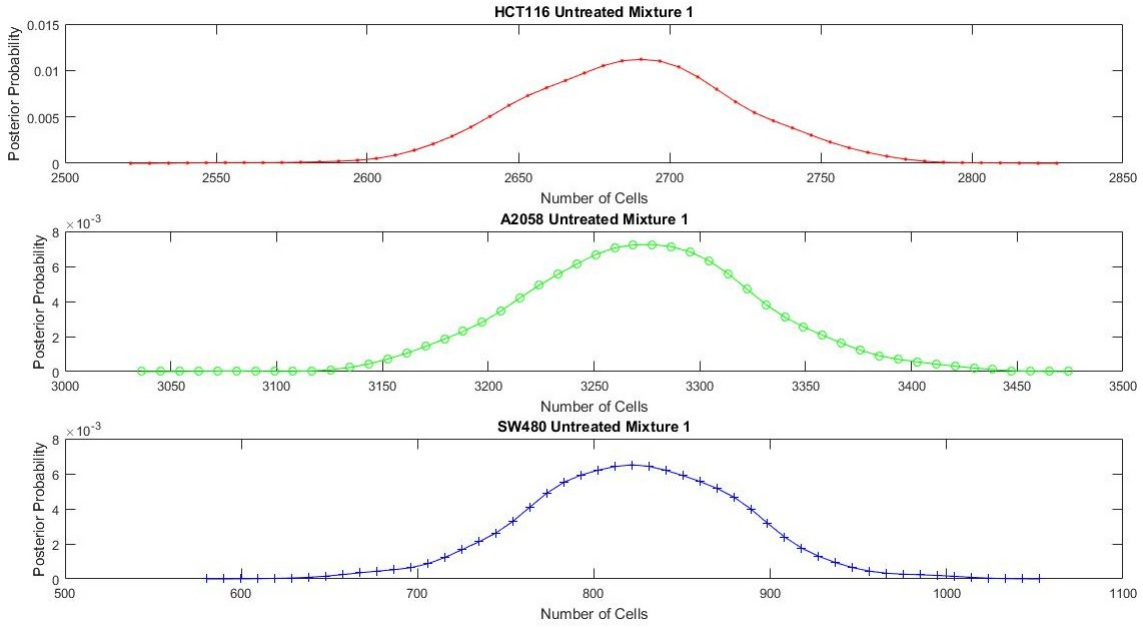

Figure 3.6: Posterior Probability Distribution of N for Untreated Mixture 1

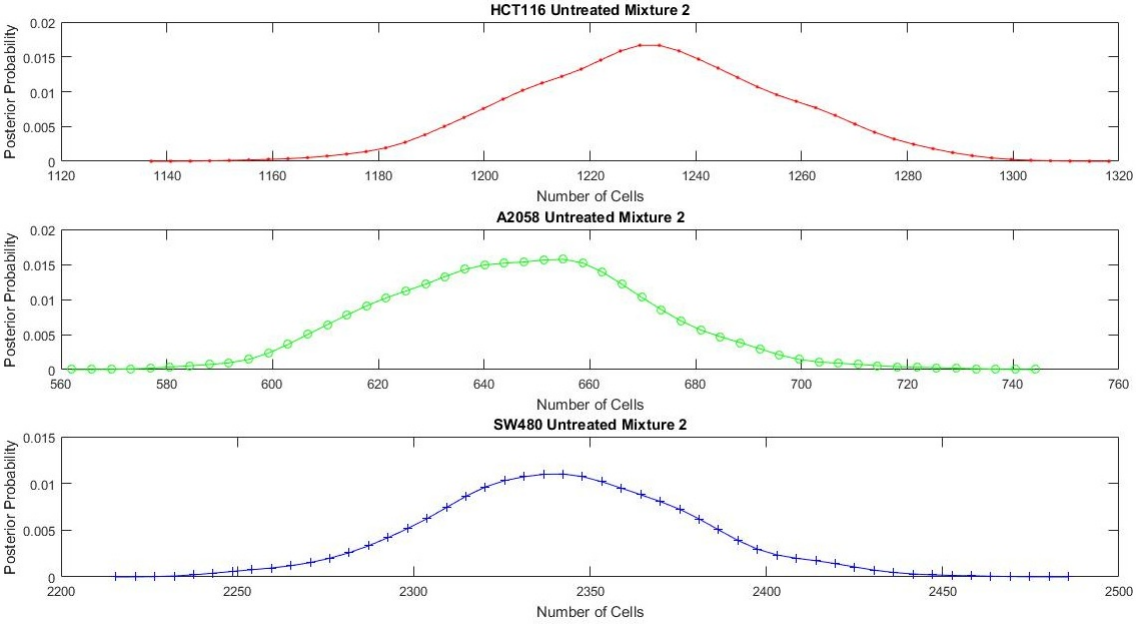

Figure 3.7: Posterior Probability Distribution of N for Untreated Mixture 2

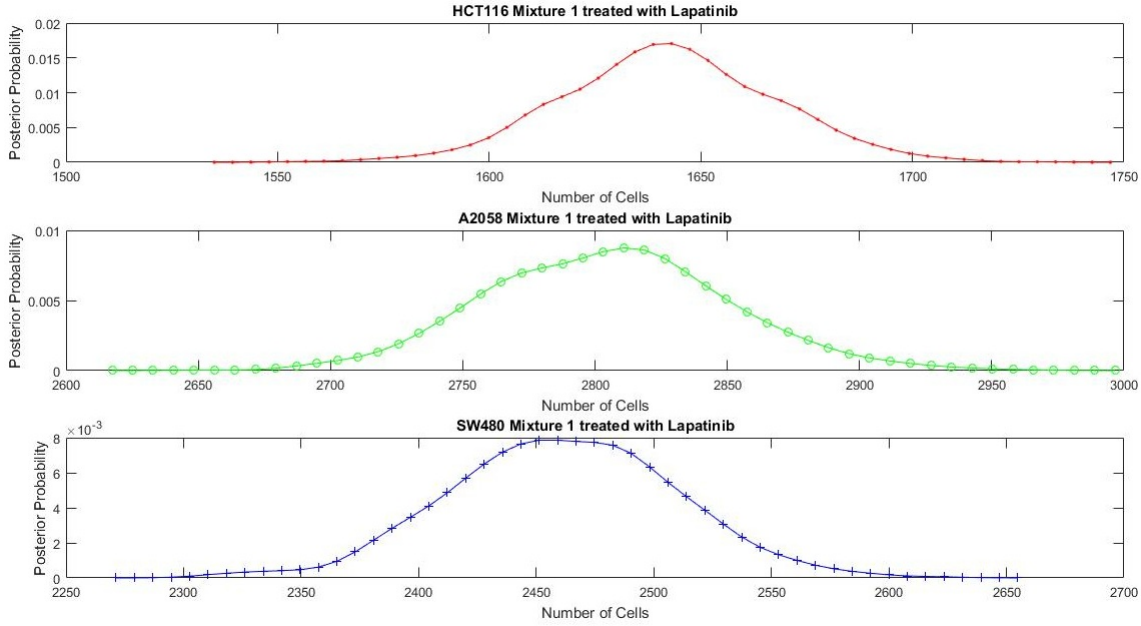

Figure 3.8: Posterior Probability Distribution of N for Mixture 1 Treated with Lapatinib

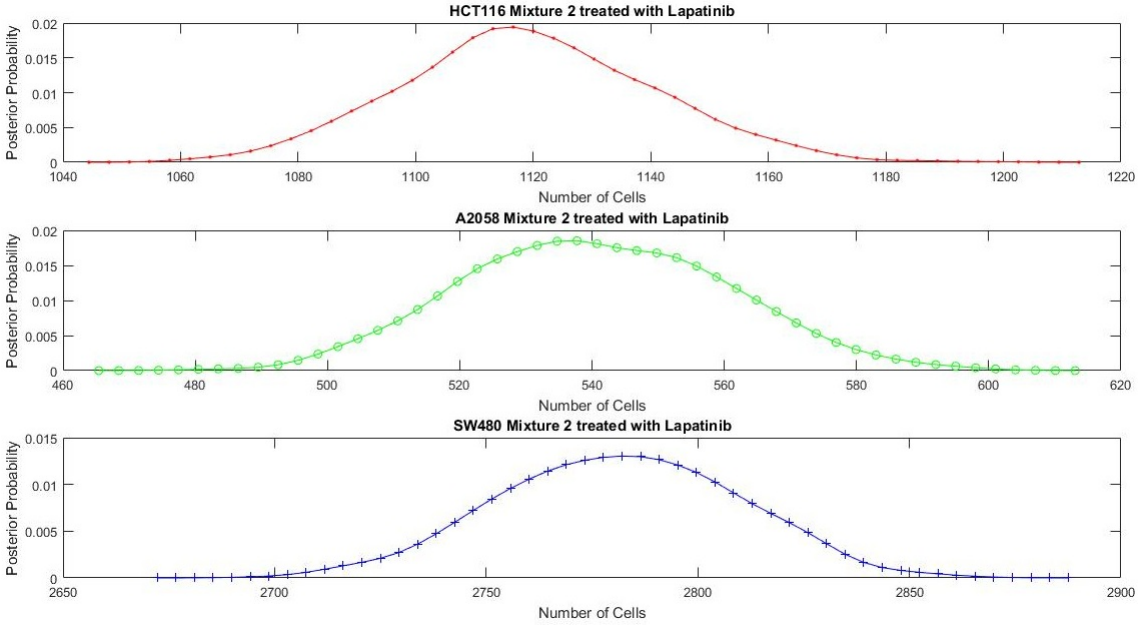

Figure 3.9: Posterior Probability Distribution of N for Mixture 2 Treated with Lapatinib

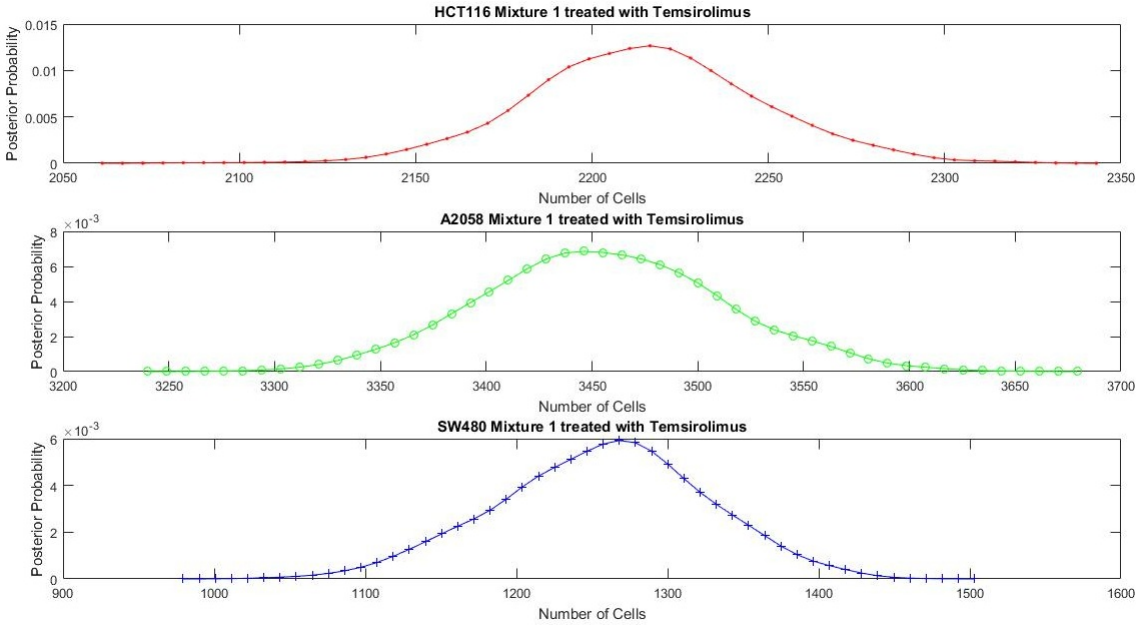

Figure 3.10: Posterior Probability Distribution of N for Mixture 1 Treated with Temsirolimus

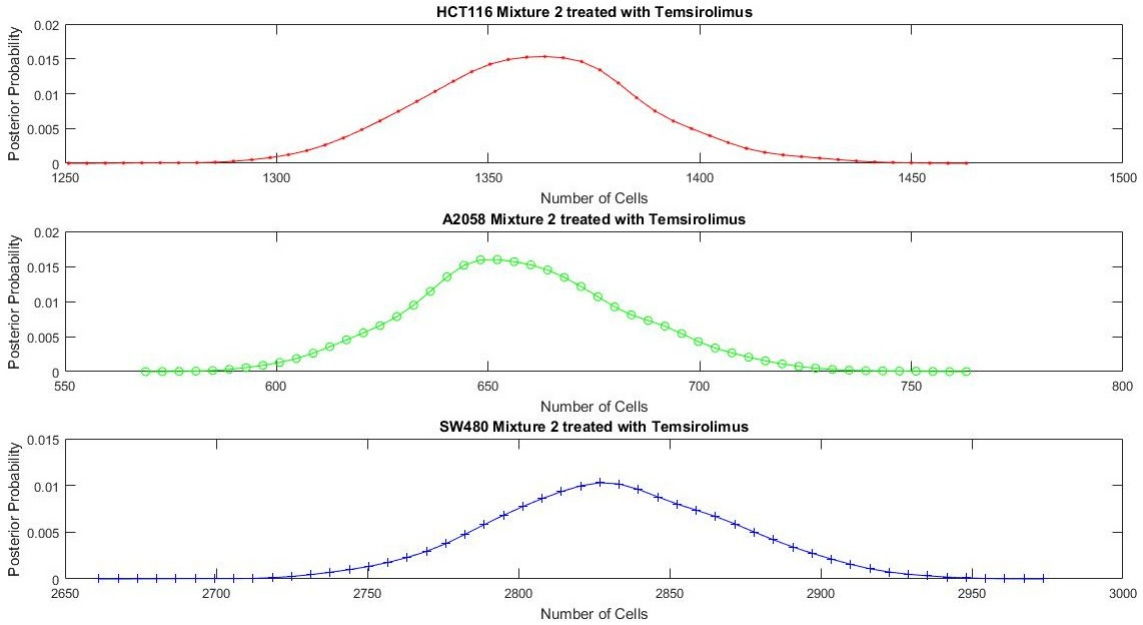

Figure 3.11: Posterior Probability Distribution of N for Mixture 2 Treated with Temsirolimus

when the number of cells of cell lines are large, especially for cell lines whose elements of response vector have high standard deviation. This is because, even though the distribution of the summation of the different components of response vector approximates to a Gaussian distribution, the overall response can be far from the peak of the Gaussian if there are not enough samples.

4. SUMMARY AND CONCLUSIONS

In this work we address the challenge of determining the composition of any heterogeneous cancer tissue. It uses the advantage offered by the expensive cell-by-cell analysis methods while actually utilizing the low cost ensemble methods. The algorithm takes the characteristics of the response vector individual cell lines and the output of the ensemble method as inputs. Based on these inputs, the algorithm uses a Bayesian approach to estimate the number of cells of different cell lines that are present in the heterogeneous mixture. In order to estimate the posterior probability, the algorithm uses the Metropolis algorithm to gather samples from the posterior distribution and Kernel Density Estimation to estimate the distribution from these samples.

REFERENCES

[1] N. A. Saunders, F. Simpson, E. W. Thompson, M. M. Hill, L. EndoMunoz, G. Leggatt, R. F. Minchin, and A. Guminski, "Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives," *EMBO Mol Med*, vol. 4, no. 8, pp. 675–684, 2012.

[2] P. C. Nowel, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[3] A. K. Mohanty, A. Datta, and V. Venkatraj, "A model for cancer tissue heterogeneity," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 966–974, 1976.

[4] C. Sima, J. Hua, R. Lopes, A. Datta, and M. L. Bittner, "Detecting cell growth and drug response in heterogeneous populations: A dynamic imaging approach," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 121–128, Oct 2016.